Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke

ISLA, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Net
{m.tsagkias,w.weerkamp,derijke}@uva.nl

Abstract. Online news agents provide commenting facilities for their reader express their opinions or sentiments with regards to news stories. The number user supplied comments on a news article may be indicative of its importan interestingness, or impact. We explore the news comments space, and comp the log-normal and the negative binomial distributions for modeling comme from various news agents. These estimated models can be used to normalize comment counts and enable comparison across different news sites. We also amine the feasibility of online prediction of the number of comments, based the volume observed shortly after publication. We report on solid performan for predicting news comment volume in the long run, after short observati This prediction can be useful for identifying news stories with the potentia "take off," and can be used to support front page optimization for news sites.

1 Introduction

As we increasingly live our lives online, huge amounts of content are being and stored in new data types like blogs, discussion forums, mailing lists, co facilities, and wikis. In this environment of new data types, online news is an interesting type for mining and analysis purposes. Much of what goes on in dia is a response to, or comment on, news events, reflected by the large news-related queries users ask to blog search engines [9]. Tracking news e their impact as reflected in social media has become an important activity of alysts [1] and there is a growing body of research on developing algorithms to support this type of analysis (see the related work section below). In this focus on online news articles plus the comments they generate, and attempt t the factors underlying the commenting behavior on these news articles. We e dynamics of user generated comments on news articles, and undertake the ch model and predict news article comment volume shortly after publication.

To make things more tangible, consider a striking example of unexpected ing behavior in response to news stories: March 13, 2009, a busy day for biggest news papers in the Netherlands, *De Telegraaf*. In less than 24 hours, 1,500 people commented on *Telegraaf*'s article regarding the latest governme on child benefit abuse. One month later, the Dutch news reported a potential swine flu, first located in Mexico, but less than five hundred comments were related articles across different news sites, even a week after the first publicati that both news events are important to the Dutch society, their numbers of

C. Gurrin et al. (Eds.): ECIR 2010, LNCS 5993, pp. 191–203, 2010.

and the factors contributing to it in the first place? We briefly mention tw application for predicting the number of comments shortly after publicatio *reputation analysis* one should be able to quickly respond to stories that "tak real-time observation and prediction of the impact of news articles is require the *lay-out decisions* of online news agents often depend on the expected articles, giving more emphasis to articles that are likely to generate more c both in their online news papers (e.g., larger headline, picture included) an RSS feeds (e.g., placed on top, capitalized).

To come to these applications and answer the questions raised by the ex need more insight in comments and commenting behavior on online news art aim is to gain this insight, and use these insights to predict comment volum articles shortly after publication. To this end, we seek to answer the following

1. What are the dynamics of user generated comments on news articles follow a temporal cycle? The answers provide useful features for mo predicting news comments.
2. Can we fit a distribution model on the volume of news comments? Mo distribution allows for normalizing comment counts across diverse news
3. Does the correlation between number of responses at early time and at found in social media such as Digg and Youtube hold for news comment patterns for online responses potentially "universal"? And can we use this the number of comments an article will receive, having seen an initial nu

This paper makes several contributions. First, it explores the dynamics and the cycles of user generated comments in online Dutch media. Second, it provide for news comment distribution based on data analysis from eight news sou third, it tries to predict comment volume once an initial number of comments using a linear model. In §2 we discuss related work. §3 explores the datase use insights gained here to try to fit distribution models in §4. Finally, we try comment volume in §5 and conclude in §6.

2 Related Work

Different aspects of the comment space dynamics have been explored in Mishne and Glance [10] looked at weblog comments and revealed their usef improving retrieval and for identifying blog post controversy. Duarte et al. [4 in describing blogosphere access patterns from the blog server point, and three groups of blogs using the ratio of posts over comments. Kaltenbrunne measured community response time in terms of comment activity on Slashd and discovered regular temporal patterns in people's commenting behavio Salamatian [7] report that the amount of comments in a discussion thread is proportional to its lifespan after experimenting with clustering threads for t discussion fora, and for a social networking site. Schuth et al. [12] explore

We explore the comment space of online news articles, and model the co
patterns for multiple news sources. Previous work finds that the distributio
ments over blog posts is governed by Zipf's law [8, 10, 12]. Lee and Salar
use a Weibull distribution for modeling comments in discussion threads. Kalt
et al. [5] point to discussions in the literature for selecting the log-normal ove
distribution for modeling; they use four log-normal variants to model respons
Slashdot stories. Ogilvie [11] models the distribution of comment counts in
using the negative binomial distribution; a similar approach is taken by Tsag
[15] to model news comments for prediction prior to publication. Finally, W
berman [16] find that diggs can be modeled with the log-normal distribution,
and Huberman [14] model popularity growth of online content using a linear

3 Exploring News Comments

In this section we describe our data, comments to online news articles, com
menting behavior to that in the blogosphere, and discover temporal cycles.

The dataset consists of aggregated content from seven online news age
meen Dagblad (*AD*), *De Pers*, *Financieel Dagblad* (*FD*), *Spits*, *Telegraaf*, *T*
WaarMaarRaar (*WMR*), and one collaborative news platform, *NUjij*. We
sen to include sources that provide commenting facilities for news stories,
in coverage, political views, subject, and type. Six of the selected news ag
lish daily newspapers and two, *WMR* and *NUjij*, are present only on the w
publishes "oddly-enough" news and *NUjij* is a collaborative news platform,
Digg, where people submit links to news stories for others to vote for or initi
sion. We focus only on the user interaction reflected by user generated com
other interaction features may play a role on a user's decision to leave a com

For the period November 2008–April 2009 we collected news articles and
ments. Our dataset consists of 290,375 articles, and 1,894,925 comments. T
is mainly written in Dutch. However, since our approach is language indepe
we believe that the observed patterns and lessons learned apply to news co
other countries, we could apply our approach to other languages as well.

3.1 News Comments vs. Blog Post Comments

The commenting feature in online news is inspired by the possibility for blog
leave behind their comments. Here, we look at general statistics of our news s
comments, and compare these to commenting statistics in blogs as reported i
numerical summary can be found in Table 1. News comments are found to fol
similar to blog post comments. The total number of comments is an order of
larger than the total number of articles, which is positively correlated with t
influential blogs. In general, about 15% of the blog posts in the dataset in [1
comments, a number that increases for the news domain: the average perc
commented articles across all sources in our dataset is 23%. *Spits* and *WMR*

News agent	Total articles (commented)		Total comments	Comments per article w/ comments			Time (h
				mean	median	st.dev	0–1 com. 1–l
AD	41 740	(40%)	90 084	5.5	3	5.0	9.4
De Pers	61 079	(27%)	80 72	5.0	2	7.5	5.9
FD	9 911	(15%)	4 413	3.0	2	3.8	10.
NUjij	94 983	(43%)	602 144	14.7	5	32.3	3.1
Spits	9 281	(96%)	427 268	47.7	38	44.7	1.1
Telegraaaf	40 287	(21%)	584 191	69.9	37	101.6	2.5
Trouw	30 652	(8%)	19 339	7.9	4	10.3	11.7
WMR	2 442	(100%)	86 762	35.6	34	13.08	1.1

interesting characteristic of receiving comments on almost every article the This can be explained by the two sites having very simple commenting fa contrast, *Trouw* has the lowest ratio of commented articles: commenting i only for some of the articles, partially explaining the low ratio of commente Another reason can be the content's nature: WMR's oddly-enough news item accessible and require less understanding increasing the chance to be comme

Half of the news sources receive the same number of comments as blogs (whereas the other half enjoys an order of magnitude more comments than blo ing at reaction time, the time required for readers to leave a comment, it is c slower for news ($\sim$ 6 hrs) than for blogs ($\sim$ 2 hrs), although this differs sig per news source. A speculation on the reason underlying these differences c news source's readers demographics, e.g., tech savvies or youngsters are ra to react, whilst older people, less acquainted with the internet, access the onli of the news papers less frequently.

3.2 Temporal Cycles of News Comments

We perform an exploration of temporal cycles governing the news comment look at three levels of temporal granularity: monthly, daily, and hourly. In o

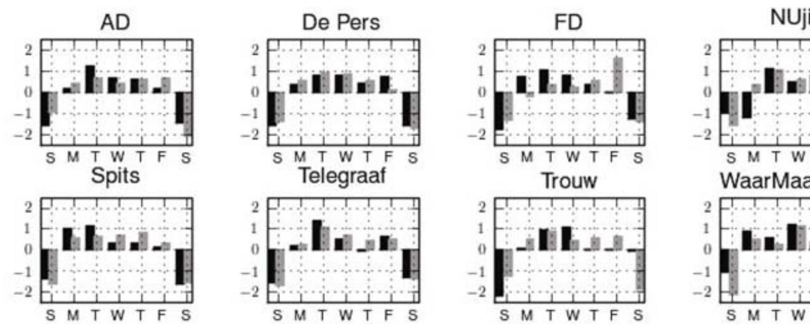

Fig. 1. Comments (black) and articles (grey) per day of the week and per sou

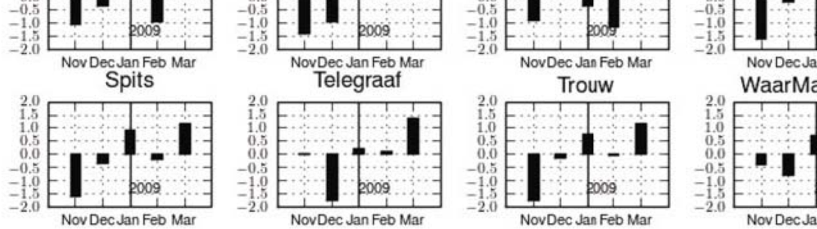

Fig. 2. Comments per month and per source. Vertical line is a year separator

the volume of comments ranges two orders of magnitude making the comparis
comment counts difficult. We therefore report comments in z-scores: z-scores
how many σ's (standard deviations) the score differs from the mean, and a
comparison across sources.

Looking at comment volume per month in Fig. 2, we observe months
and low comment volume, either reflecting the importance of published ne
seasonal user behavior. For example, March shows the highest comment volu
the board, and November shows the least for most sources.

We explore the comment volume per day of the week in Fig. 1: weekda
more comments compared to weekends, with Wednesday being, on average
active day and Sunday the least active day across the board. These results are
ment with the activity observed in social networks such as Delicious, Digg, an
Comparing the number of comments to the number of articles published per
sources show an insignificant, negative correlation ($p \gg 0.05$). Three sources
have articles and comments highly correlated, but differ in polarity: *FD* and *Tr*
a negative correlation and *NUjij* shows a positive correlation. The variety fou
relation polarity likely indicate the commenting behavior of a source's audie

Finally, we look at the distribution of comments throughout the day. Fig.
a correlation between posted comments, sleep and awake time, as well as
lunch and dinner time. The comment volume peaks around noon, starts dec
the afternoon, and becomes minimal late at night. Interesting exceptions are
collaborative news platform, and *FD*, a financial newspaper: comment volum
matches with blog post publishing [8], which has a slow start and gradually pe
the evening. *FD* on the other hand receives most of its comments early in the
and then drops quickly. This is in line with the business oriented audience of

Overall, the commenting statistics in online news sources show similaritie
in the blogosphere, but are nevertheless inherent characteristics of each new
The same goes for the temporal cycles, where we see similar patterns for mos
but also striking differences. These differences in general and temporal char

[1] `http://3.rdrail.net/blog/thurday-at-noon-is-the-best-ti`
`-and-be-noticed-pst/`

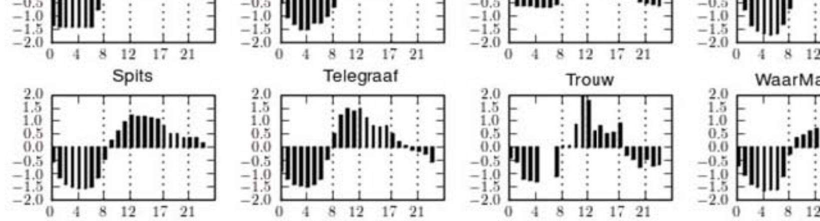

Fig. 3. Comments per hour and per source

possibly reflect the credibility of the news organisation, the interactive fea
provide on their web sites, and their readers' demographics [2].

4 Modeling News Comments

In this section we seek to identify models (i.e., distributions) that underly th of comments per news source. We do so (1) to understand our data, and (2) "volume" across sources. If two articles from two sources receive the sam of comments, do they expose the same volume? Ten comments may sign volume for an article in one source, but a low volume in another. Expressing volume as a normalized score offers a common ground for comparing and articles between sources. Our approach is to express a news article's comme as the probability for an article from a news source to receive x many comm consider two types of distribution to model comment volume: log-normal an binomial.

Recall that the log-normal distribution is a continuous distribution, with ity density function defined for $x > 0$, cf. (1), and the two parameters μ (and σ (the standard deviation of the variable's natural logarithm) affect th tion's shape. For a given source we estimate the parameters using maximum estimation.

$$LN_{pdf}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln x - \mu)^2}{2\sigma^2}}$$

The negative binomial distribution is a discrete distribution with probability n tion defined for $x \geq 0$, with two parameters r ($r - 1$ is the number of times a occurs) and p (the probability of observing the desired outcome), cf. (2). T analytical solution for estimating p and r, but they can be estimated numerica

$$BN_{pmf}(k; r, p) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k$$

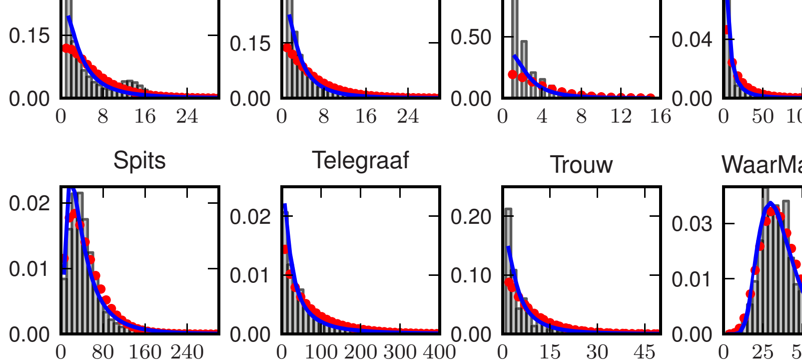

Fig. 4. Modeling comment volume distribution per source using the continuous log-no line), and the discrete negative binomial distribution (red dots). Grey bars represen data. Probability density is on y-axis, and number of comments (binned) is on x-axis.

For evaluating the models' goodness of fit we choose the χ^2 test. χ^2 is a good a to the widely used Kolmogorov-Smirnov goodness of fit test due to its appli both continuous and discrete distributions [13]. The metric tests whether a observed data belongs to a population with a specific distribution. Note tha requires binned data, and as such is sensitive to the number of chosen bins.

For each news source we estimate the parameters for the log-normal and th binomial distributions over the entire period of our dataset (see Fig. 4), and goodness of fit results in Table 2. Both distributions fit our dataset well, wi scores denoting strong belief that the underlying distribution of the data matc log-normal and negative binomial. Log-normal is rejected for *WaarMaarRaa* because it failed to reach close enough the peak observed at 25 comments. that the results should be taken as indicative, mainly due to the sensitivity of number of bins (here 30). We experimented with different bin sizes, and obs for different number of bins either the log-normal, or the negative-binomia describe all sources. Although searching for the optimal number of bins for tributions to fit all sources could be interesting, we did not exhaust the entire An example of the test's sensitivity is shown in Table 3 where log-normal dis similar results to negative-binomial even for the source that failed the χ^2 test

The final decision on which distribution to favor, depends on the data to be and task at hand. From a theoretical point of view, negative binomial seem be to the task of modeling comments: comments are not a continuous but disc able. From a practical point of view, for the same task, log-normal paramete expensive to estimate and the results match closely those of negative binomia

The results of our data exploration and modeling efforts are put to the next section, in which we explore the correlation between comment volume s longer after publication.

proved to be very challenging [15]. Szabó and Huberman [14] published
work on predicting the long term popularity of Digg stories (measured in d
Youtube videos (measured in views) *after* observing how their popularity evo
first hours of publication. First, we are interested in finding out whether the c
between early and late popularity found by Szabó and Huberman also hol
news comments space. Then, assuming such a relation has been confirmed
employed for predicting the comment volume of a news story.

We begin to explore the relation between early and late comment volum
ing at the similarities of news comments and other online user generated c
Section 3.2 we reported on the circadian pattern underlying news comment g
which is found to be similar to blog posts [10], Diggs and Youtube video v
The existence of a circadian pattern implies that a story's comment volum
on the publication time, and therefore not all stories share the same prospec
commented; stories published during daytime—when people comment the m
a higher prior probability of receiving a comment.

Taking into account the above, publication time adds another dimensio
plexity in finding temporal correlations. To simplify our task, we introduce a
transformation from real-time to *source-time*, following [14], a function of
ment volume entering a news site within a certain time unit. I.e., *source-time*
as the time required for $\overline{x}_i$ comments to enter a news agent system i, where
for the average comments per hour cast to a particular source, and is the d
a source's total comments by the total number of hours that we have data fo
quently, *source-time* has the property of expanding or condensing the real-t
in order to keep the ratio of incoming comments per hour fixed. Once the n
comments per time unit has been fixed, all stories share the same probab
commented independently of their publication time. In the rest of this sect
comments are measured in their news agent specific *source-time*, e.g., for
measure in *trouw-time*, for *WMR* in *wmr-time*, etc. Once the temporal transfo

Table 2. χ^2 goodness of fit for log-normal and negative binomial distributions at 0.10 s
level. Boldface indicates rejection of the null hypothesis: observed and expected dat
the same distribution.

News site	Log-normal		Negative binomial	
	χ^2 score	p-value	χ^2 score	p-value
AD	0.08	1.00	0.08	1.00
De Pers	0.59	1.00	0.64	1.00
FD	0.18	1.00	0.26	1.00
NUjij	0.06	1.00	0.06	1.00
Spits	0.67	1.00	1.42	1.00
Telegraaf	0.04	1.00	0.04	1.00
Trouw	0.56	1.00	0.98	1.00
WaarMaarRaar	**236.89**	0.00	0.15	1.00

Distribution	Comments for ICDF @ 0.5							
	AD	*De Pers*	*FD*	*NUjij*	*Spits*	*Telegraaf*	*Trouw*	*WMI*
Log-normal (LN)	3	3	2	6	36	32	4	34
Negative binomial (NB)	3	3	1	8	39	43	5	33

in place, we need a definition for *early* and *late* time, between which we are in discovering a correlation. We introduce *reference time* t_r as "late" time, a it at 30 source-days after the story has been published. For "early" time, we *dicator time* t_r to range from 0 to t_r in hourly intervals [14]. Some news ager comments after a certain period. As a result, there are articles that constantly r maximum comments before t_r, however we have not marked them separately

We choose Pearson's correlation coefficient ρ to measure the correlation st tween reference and indicator times. Using articles with more than one com compute ρ in hourly intervals from publication time to reference time for a over the entire period of the dataset. Fig. 5 shows that the number of com source increases exponentially, yet with different rates, reflecting the commen of each site: the time a story remains visible on the front page, for how long are enabled, etc. In the same figure we show a positive correlation that grow as t_i approaches t_r due to stories that saturate to their maximum number of c The curve slope indicates how fast stories reach their maximum number of c e.g., *Spits* displays a very steep comment volume curve meaning that most st receiving comments short after publication. Looking at when sources reach s relation ($\rho > 0.9$) we find that the corresponding indicator times reflect th

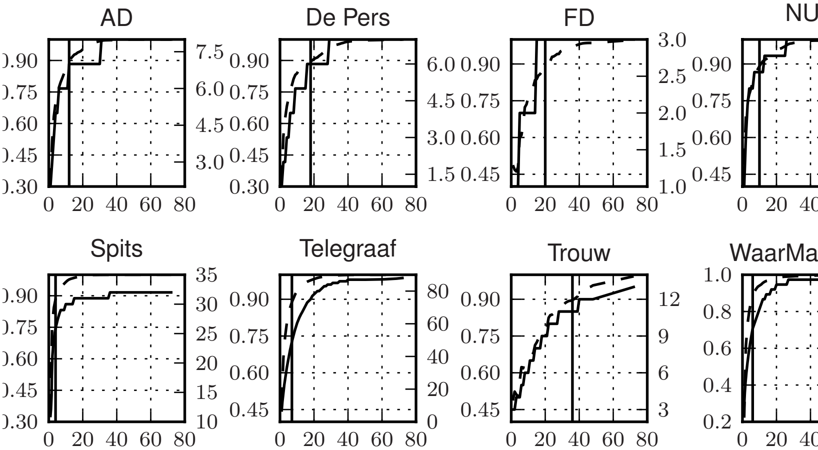

Fig. 5. Comment counts averaged over all stories (right y-axis, solid line), and ρ bet cator, and reference time (left y-axis, dashed line). Indicator time shown at x-axis. V shows the indicator time with $\rho \geq 0.9$.

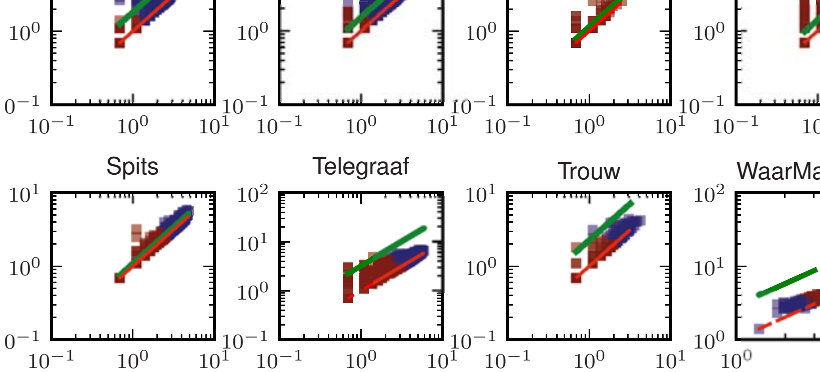

Fig. 6. Correlation of news stories comment volume per source between 2 hours, a
after publication. Number of comments at $t_i(2)$ is x-axis, and comments at t_r is y-axi
separates stories in two clusters depending on their initial comments. Green line sho
model using only the upper stories, with slope fixed at 1. Red dashed line marks th
where no stories can fall below.

commenting lifespan of each source (see Table 1). In contrast to our expecta
NUjij, the collaborative news platform, follows a fast correlation pattern simil
(0.98 after the 5th digg-hour), our findings suggest that a strong correlation is
much later (ρ at 0.90 after 11 source-hours). Although, nujij-time and digg
not directly comparable, we can compare the average user submissions ente
system per hour: 5.478 diggs vs. 140 comments. The difference in order of m
can be explained by the different popularity levels enjoyed by the two sites. C
argue that digg-ing and commenting are different tasks: on the one hand, com
similarly to writing, asks for some reflection on how to verbalize one's though
less of the size or the quality of the comment. On the other hand, digg-ing re
click of a button, rendering the task easier, and hence more attractive to partic

Given the exponential accumulation of comments over time, a logarithm
appropriate for plotting. In contrast to Diggs or YouTube views, comments do
more than two orders of magnitude (compare $10^0 - 10^2$ for comments to 10^1
Diggs and Youtube views). Despite the difference in scale, our data shows an
pattern similar to Youtube, where a bump is observed in the middle rang
comments. From Fig. 6 two groups of stories emerge, both resulting in many c
one with stories that begin with too few comments in early indicator times, and
stories that begin with many comments. This pattern is different from Digg o
where a linear correlation is evident in similar graphs [14].

For our prediction experiments, we are interested in minimizing noise to
performance, and hence could exploit the emerging clusters by eliminating st
too few comments at early indicator times. Since these stories exhibit a rathe

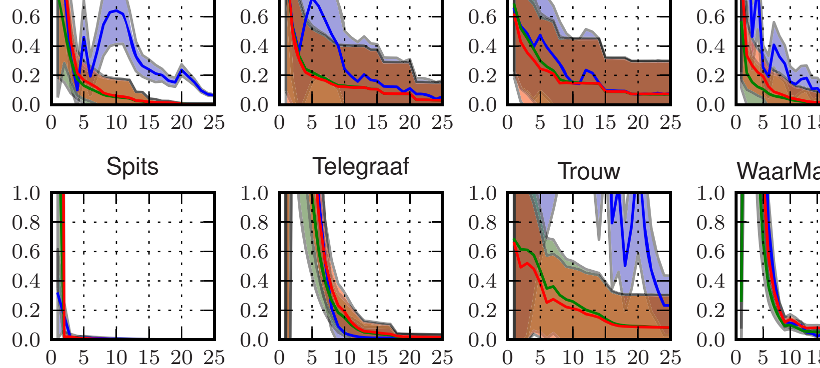

Fig. 7. Relative square error using Model 1 (blue line), Model 2 (green line), and M
line). Standard deviation is shown in the shaded areas around the lines. QRE on y-axi
time on x-axis.

pattern with regards to their final number of comments, we employ k-means
in an attempt to separate them from stories that show a more consistent patte

We follow [14] and estimate a linear model on a logarithmic scale for ea
in our dataset. The linear scale estimate $\hat{N}_s$ for a story s at indicator time t
is defined as $\hat{N}_s(t_i, t_r) = exp[ln(\alpha_0 N_s(t_i)) + \beta_0(t_i) + \sigma^2/2]$, where N_s
observed comment counts, α_0 is the slope, β_0 is the intercept, and σ^2 is the v
the residuals from the parameter estimation.

For evaluating our model we choose the relative squared error metric aver
all stories from a certain source at t_i given t_r.

$$QRE(s, t_i, t_r) = \sum_c \left[\frac{\hat{N}_s(t_i, t_r) - N_s(t_r)}{N_s(t_r)} \right]^2$$

For our experiments, we split our dataset in training and testing for each so
training sets span from November 2008—January 2009, and the test sets cove
2009. Model parameters are estimated on the training set, and QREs are calc
the test set using the fitted models.

We define three experimental conditions based on which we estimate mode
ters using our training set: (M1) using in the upper end stories as clustered by
and fixing the slope at 1, (M2) using all stories, and fixing the slope at 1, and
ing all stories. Fig. 7 illustrates QREs for the three experimental conditions
hours after observation; we choose not to include all indicator times up to
time to increase readability of the details at early times. From the three exp
conditions, M1 proved to underperform in most cases. M2 and M3 demonstra
performance across the board with one slightly outperforming the other dep
the source. QREs decrease to 0 as we move to reference time, followed by

commenting dynamics of each source as discussed earlier.

In this section we looked at natural patterns emerging from news comm as the possible correlation of comment counts on news stories between early publication time. A relation similar to the one observed for Digg and Youtub confirmed, allowing us to predict long term comment volume with very small observed that different news sources ask for different observation times befo prediction can be made. Using QRE curves one can find the optimum observ per source, that balances between short observation period and low error.

6 Conclusion and Outlook

We studied the news comments space from seven online news agents, and laborative news platform. Commenting behavior in the news comments spa similar trends as the behavior in the blogosphere. Our news sources show qu temporal cycles and commenting behavior, but that mainly the differences he differences in readers' demographics and could prove useful in future researc

As to modeling comments from various news agents, we compared the l and negative binomial distributions. These estimated models can be used to raw comment counts and enable comparison, and processing of articles fron news sites. According to χ^2 goodness of fit test, the underlying distributio comments matches with either log-normal or negative binomial. The latter is distribution and suits the task better, yet in our our setup log-normal show results and parameter estimation for log-normal is computationally less expe

Finally, we looked at the feasibility of predicting the number of comment time, based on the number of comments shortly after publication. Our goal w patterns similar to other online content such as Digg, and Youtube. We conf relation, and exploited its potential using linear models. Our results showed th tion of the long term comment volume is possible with small error after 10 sou observation. This prediction can be useful for identifying news stories with th to "take off," and can for example be used to support front page optimizatio sites.

Acknowledgments. This research was supported by the DAESO and projects carried out within the STEVIN programme which is funded by the Flemish Governments under project numbers STE-05-24 and STE-09-12, a Netherlands Organisation for Scientific Research (NWO) under project num 001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802

References

[1] Altheide, D.L.: Qualitative Media Analysis (Qualitative Research Methods). S Inc., Thousand Oaks (1996)

[2] Chung, D.S.: Interactive features of online newspapers: Identifying patterns and use of engaged readers. J. Computer-Mediated Communication 13(3), 658–679

communication patterns in blogosphere. In: ICWSM 2006 (March 2007)
[5] Kaltenbrunner, A., Gomez, V., Lopez, V.: Description and prediction of slasho
In: LA-WEB 2007, pp. 57–66 (2007)
[6] Kaltenbrunner, A., Gómez, V., Moghnieh, A., Meza, R., Blat, J., López, V.: Ho
temporal activity patterns in a large online communication space. CoRR (2007)
[7] Lee, J.G., Salamatian, K.: Understanding the characteristics of online comm
CONEXT 2008, pp. 1–2 (2008)
[8] Mishne, G., de Rijke, M.: Capturing global mood levels using blog posts. In: AA
pp. 145–152 (2006)
[9] Mishne, G., de Rijke, M.: A study of blog search. In: Lalmas, M., MacFarlane,
S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol
289–301. Springer, Heidelberg (2006)
[10] Mishne, G., Glance, N.: Leave a reply: An analysis of weblog comments. In: T
workshop on the Weblogging ecosystem (2006)
[11] Ogilvie, P.: Modeling blog post comment counts (July 2008),
http://livewebir.com/blog/2008/07/
modeling-blog-post-comment-counts/
[12] Schuth, A., Marx, M., de Rijke, M.: Extracting the discussion structure in co
news-articles. In: WIDM 2007, pp. 97–104 (2007)
[13] Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedure
& Hall/CRC, USA (2000)
[14] Szabó, G., Huberman, B.A.: Predicting the popularity of online conte
abs/0811.0405 (2008)
[15] Tsagkias, M., Weerkamp, W., de Rijke, M.: Predicting the Volume of Comment
News Stories. In: CIKM 2009, pp. 1765–1768 (2009)
[16] Wu, F., Huberman, B.A.: Novelty and collective attention pnas,
http://www.pnas.org/content/104/45/17599.abstract